Bayesian Decision Theory

- Bayesian Classification
- Classification Error
- Losses and Risks
- Discriminant Functions
- The Normal Density
- Bayesian Classification for Normal Distribution
- Naïve-Bayes Classifier

Probability and Inference

- Making inference from data
  - The data generating process maybe deterministic
    - $x = f(z)$
      - z: the unobservable variable
      - x: the observable variable (e.g., outcome of an experiment)
    - But we do not have access to the complete knowledge of $f(.)$?
  - We model the process as random
    - By defining the outcome X as a random variable drawn from $P(X = x)$
- Bayes' rule (check Appendix A for basic probability theory)
  - When 2 random variables X, Y are jointly distributed
    - With the value of one known $X = x$, the probability that the other takes a given value $Y = y$ can be calculated by $P(y|x) = \frac{P(x|y)P(y)}{P(x)}$

Bayesian Classification

- Bayesian classification
  - Assumption
    - Quantities of interest are governed by probability distributions
      - Statistical variations of the generated features
  - To model and quantify our uncertainty for hypotheses
    - By combining prior knowledge and observed data
    - Accommodate hypotheses that make probabilistic predictions
      - e.g., a patient has a 90% chance of recovery
    - $P(C|X) = \frac{P(X|C)P(C)}{P(X)} \Rightarrow \text{posterior} = \frac{\text{likelihood}\times\text{prior}}{\text{evidence}}$
  - Determine the **best** hypothesis as
    - The **most probable one** given the observed data

Bayes' Theorem

- Bayes' Theorem
  - Let $C_i,\ i = 1, 2, \dots, K$ be a set of disjoint events with $P(C_i) > 0$
  - For any event X with $P(X) > 0$
    - $P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} = \frac{P(X|C_i)P(C_i)}{\sum_{j=1}^{K} P(X|C_j)P(C_j)}$
    - $C_1, \dots, C_K$: hypotheses
    - $P(C_i)$: the prior probability of C_i
    - $P(C_i|X)$: the posterior probability of C_i after the occurrence of X

Law of total probability

$$P(X) = \sum_{j=1}^{K} P(X|C_j)P(C_j)$$

Reasoning from the data to hypotheses (inverse reasoning) is often much more difficult than reasoning from the hypothesis to the data (forward reasoning)

  - To calculate $P(C_i|X)$
    - P(the hypothesis C_i given the observed data X)

    $\propto P$(the observed data X given the hypothesis C_i) $\times P(C_i)$

Example (1)

- The sea bass & salmon classifier [Duda 01]
  - State of nature C
    - C is considered as a random variable
      - $C = C_1$ for sea bass
      - $C = C_2$ for salmon
  - Prior (a priori probability): $P(C_1), P(C_2)$
    - $P(C_1) + P(C_2) = 1$
    - Prior knowledge of how likely we are to get a sea bass or salmon before the fish actually appears
      - May depend on the time of year or the choice of fishing area
  - Decision with only the prior information
    - $\mathbf{x} \to C_1$ if $P(C_1) > P(C_2)$; $\mathbf{x} \to C_2$ otherwise
    - Error rate = $\min\{P(C_1), P(C_2)\}$

Always make the same decision for all the fish caught

Example (2)

- The sea bass & salmon classifier (cont.)
  - Class-conditional probability density function: $p(x|C_1), p(x|C_2)$
    - If having the lightness measurement x from the two kinds of fish

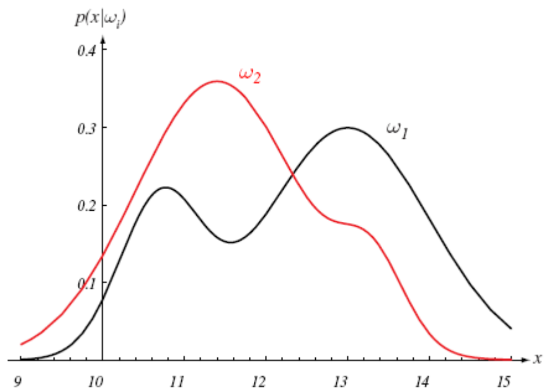


Fig. 2.1 [Duda 01]

  - Maximum likelihood decision rule
    - $x \to C_1$ if $p(x|C_1) > p(x|C_2)$; $x \to C_2$ otherwise
    - The category C_i with larger $p(x|C_i)$ is more likely to be the true one

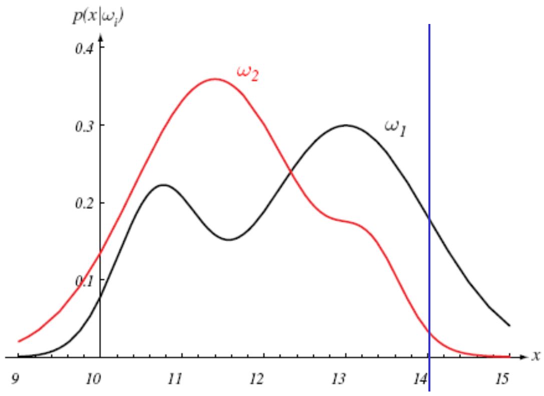
Example (3)

- The sea bass & salmon classifier (cont.)
  - Suppose now we have
    - The prior probabilities $P(C_1), P(C_2)$
    - The conditional densities $p(x|C_1), p(x|C_2)$
  - Suppose we measure the lightness of a fish with the value x
    - How does this measurement influence our prior concerning the true state of nature?
  - Posterior (a posteriori probability): $P(C_1|x), P(C_2|x)$
    - $P(C_i|x) = \frac{p(x|C_i)P(C_i)}{p(x)} = \frac{p(x|C_i)P(C_i)}{\sum_{j=1}^{2} p(x|C_j)P(C_j)}$
  - Bayes decision rule
    - $x \rightarrow C_1$ if $P(C_1|x) > P(C_2|x)$; $x \rightarrow C_2$ otherwise
      - Decision based on the posterior probabilities

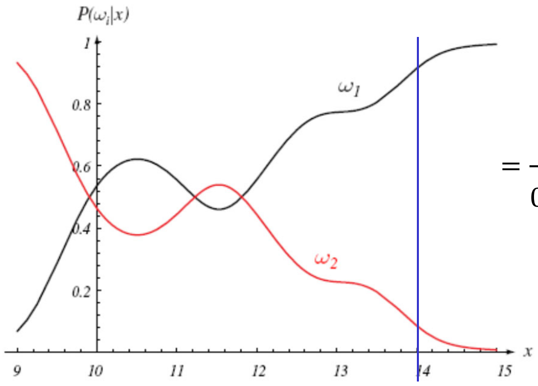
Example (4)

- The sea bass & salmon classifier (cont.)
  - Priors: $P(C_1) = \frac{2}{3}, P(C_2) = \frac{1}{3}$
  - A pattern with the feature value $x = 14$

$p(x|C_i)$



$P(C_i|x)$



$$P(C_1|x=14) = \frac{0.1725 \times \frac{2}{3}}{0.1725 \times \frac{2}{3} + 0.03 \times \frac{1}{3}} = 0.92$$

$$P(C_2|x=14) = \frac{0.03 \times \frac{1}{3}}{0.1725 \times \frac{2}{3} + 0.03 \times \frac{1}{3}} = 0.08$$

$$P(C_1|x) + P(C_2|x) = 1$$

$x = 14 \rightarrow C_1$ because $P(C_1|x) > P(C_2|x)$

Bayes Decision Theory (1)

- To classify a pattern to its most probable class
  - In a classification task of K classes
    - $\{C_1, \dots, C_K\}$
  - The unknown pattern is represented by a feature vector
    - $\mathbf{x} = [x_1, \dots, x_l]^T$
  - The K conditional probabilities
    - The *a posteriori* (or *posterior*) *probabilities*
      - $P(C_i|\mathbf{x}), i = 1, \dots, K$
    - The probability that the unknown pattern belongs to the class C_i, given that the feature vector $\mathbf{x}$ has been observed

- The minimum error classifier
  - $\mathbf{x} \rightarrow C_i$ if $P(C_i|\mathbf{x}) > P\left(C_j\middle|\mathbf{x}\right), \text{for all } j \neq i$

Bayes Decision Theory (2)

- Computation of the posterior probability
  - $P(C_i|\mathbf{x}) = \frac{p(\mathbf{x}|C_i)P(C_i)}{p(\mathbf{x})}$ $\quad\quad$ $\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$
  - The *a priori* (prior) probabilities $P(C_1), \dots, P(C_K)$
    - Usually assumed to be known
    - If not, they can be estimated from the training patterns
      - $P(C_i) \approx N_i/N$
        - » $N = \sum_i N_i$; N_i: number of training patterns belonging to C_i
  - The class-conditional probability density functions
    - $p(\mathbf{x}|C_1), \dots, p(\mathbf{x}|C_K)$
    - Also called the likelihood function of C_i with respect to $\mathbf{x}$
    - Describe the distribution of feature vectors $\mathbf{x}$ in each of the classes
    - If unknown, these functions can be estimated from training patterns

Bayes Decision Theory (3)

- Two-category case
  - $\mathbf{x} \to C_1$ if $P(C_1|\mathbf{x}) > P(C_2|\mathbf{x}); \mathbf{x} \to C_2$ otherwise
  - By eliminating the scale factor
    - The equivalent decision rule
    - $\mathbf{x} \to C_1$ if $p(\mathbf{x}|C_1)P(C_1) > p(\mathbf{x}|C_2)P(C_2); \mathbf{x} \to C_2$ otherwise
  - The probability of classification error
    - $P(error|\mathbf{x}) = \begin{cases} P(C_1|\mathbf{x}), & \text{if } \mathbf{x} \to C_2 \\ P(C_2|\mathbf{x}), & \text{if } \mathbf{x} \to C_1 \end{cases} = \min\{P(C_1|\mathbf{x}), P(C_2|\mathbf{x})\}$
    - $P(error) = \int P(error|\mathbf{x})p(\mathbf{x})d\mathbf{x}$

Example 2 (1)

- The disease classifier [Kelleher et al., 2015]
  - Given the training dataset
    - 10 patients & 3 descriptive features

ID	HEADACHE	FEVER	VOMITING	MENINGITIS
1	true	true	false	false
2	false	true	false	false
3	true	false	true	false
4	true	false	true	false
5	false	true	false	true
6	true	false	true	false
7	true	false	true	false
8	true	false	true	true
9	false	true	false	false
10	true	false	true	true

  - A patient with the measured features
    - $\mathbf{x} = [headache = T, fever = F, vomiting = T]$
  - We need to compute the posterior probability
    - $P(C_M|\mathbf{x} = [T, F, T]) = \frac{P(\mathbf{x}=[T,F,T]|C_M) \times P(C_M)}{P(\boldsymbol{x}=[T,F,T])}$

Example 2 (2)

- The disease classifier (cont.)
  - Priors
    - $P(C_M) = 0.3, P(\neg C_M) = 0.7$
  - Likelihoods
    - $P(\mathbf{x} = [T, F, T] | C_M) = 2/3$
    - $P(\mathbf{x} = [T, F, T] | \neg C_M) = 4/7$
  - The posterior probabilities
    - $P(C_M | \mathbf{x} = [T, F, T]\,) = \frac{P(\mathbf{x}=[T,F,T]|C_M)\times P(C_M)}{P(\mathbf{x}=[T,F,T])} = \frac{\frac{2}{3}\times 0.3}{\frac{2}{3}\times 0.3+\frac{4}{7}\times 0.7} = \frac{1}{3}$
    - $P(\neg C_M | \mathbf{x} = [T, F, T]\,) = \frac{P(\mathbf{x}=[T,F,T]|\neg C_M)\times P(\neg C_M)}{P(\mathbf{x}=[T,F,T])} = \frac{\frac{4}{7}\times 0.7}{\frac{2}{3}\times 0.3+\frac{4}{7}\times 0.7} = \frac{2}{3}$
    - It is twice as probable that the patient does not have meningitis as it is that the patient does

Example 2 (3)

- The disease classifier (cont.)
  - If, given another patient with the measured features
    - $\mathbf{x}' = [headache = T, fever = T, vomiting = F]$
    - Likelihoods
      - $P(\mathbf{x} = [T, T, F] | C_M) = 0/3$
      - $P(\mathbf{x} = [T, T, F] | \neg C_M) = 1/7$
    - The posterior probabilities
      - $P(C_M | \mathbf{x} = [T, T, F]\,) = \frac{0\times 0.3}{0\times 0.3+\frac{1}{7}\times 0.7} = 0$
      - $P(\neg C_M | \mathbf{x} = [T, T, F]\,) = \frac{\frac{1}{7}\times 0.7}{0\times 0.3+\frac{1}{7}\times 0.7} = 1$
  - The problem
    - The dataset is not large enough to represent the diagnosis scenario
    - The model is overfitting to the training data

Classification Error Rate (1)

- Example (p.40, [Bishop 06])
  - When we observe a particular $\mathbf{x}$
    - Assume we partition the feature space into two regions R_1 and R_2
      - $\mathbf{x} \to C_1$ if $\mathbf{x} \in R_1$
      - $\mathbf{x} \to C_2$ if $\mathbf{x} \in R_2$
    - The classification error probability
      - $P(error|\mathbf{x}) = \begin{cases} P(C_1|\mathbf{x}), & \text{if } \mathbf{x} \to R_2 \\ P(C_2|\mathbf{x}), & \text{if } \mathbf{x} \to R_1 \end{cases}$
  - The unconditional error probability
    - $P(error) = \int P(error|\mathbf{x})p(\mathbf{x})d\mathbf{x}$

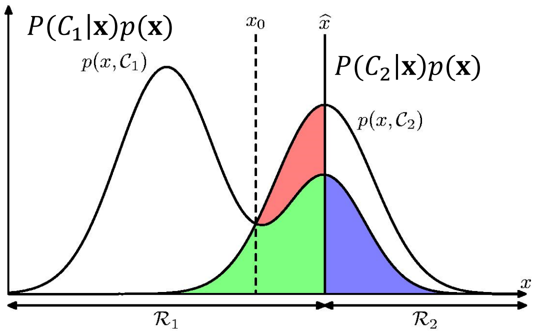


Fig. 1.24 [Bishop 06]

$$= \underbrace{\int_{R_1} P(C_2|\mathbf{x})p(\mathbf{x})d\mathbf{x}}_{\text{(red+green)}} + \underbrace{\int_{R_2} P(C_1|\mathbf{x})p(\mathbf{x})d\mathbf{x}}_{\text{(blue)}}$$

$$= P(C_1) - \int_{\mathbf{x}\to R_1} \big(P(C_1|\mathbf{x}) - P(C_2|\mathbf{x})\big)p(\mathbf{x})d\mathbf{x}$$

$$\boxed{\because P(C_1) = \int_{R_1} P(C_1|\mathbf{x})p(\mathbf{x})d\mathbf{x} + \int_{R_2} P(C_1|\mathbf{x})p(\mathbf{x})d\mathbf{x}}$$

Classification Error Rate (2)

- Error of probability of Bayesian classifier (two-category)
  - Bayes decision rule is optimal w.r.t minimizing the probability error
    - $P(error) = P(C_1) - \int_{\mathbf{x}\to R_1} \big(P(C_1|\mathbf{x}) - P(C_2|\mathbf{x})\big)p(\mathbf{x})d\mathbf{x}$
    - The the error probability is minimized if R_1 and R_2 are
      - R_1: $P(C_1|\mathbf{x}) > P(C_2|\mathbf{x})$
      - R_2: $P(C_2|\mathbf{x}) > P(C_1|\mathbf{x})$
    - Or from
      - $P(error|\mathbf{x}) = \min\{P(C_1|\mathbf{x}), P(C_2|\mathbf{x})\}$
      - $P(error) = \int P(error|\mathbf{x})p(\mathbf{x})d\mathbf{x}$

Classification Error Rate (3)

- Error of probability of Bayesian classifier (multicategory)
  - The decision rule
    - $\mathbf{x} \in R_i$ if $P(C_i|\mathbf{x}) > P(C_j|\mathbf{x})$, for all $j \neq i$
    - The probability of correct classification is maximized
    - Because R_i is chosen so that in each region the corresponding integrals have the maximum possible value
      - $P(correct) = \int P(correct|\mathbf{x})p(\mathbf{x})d\mathbf{x} = \sum_{i=1}^{K} \int_{R_i} P(C_i|\mathbf{x})p(\mathbf{x})d\mathbf{x}$
    - Thus also minimize the probability error $P(error)$
      - $\because P(error) + P(correct) = 1$

The probability error is not always the best criterion for minimization
=> Because the same importance is assigned to all errors

Losses and Risks (1)

- The penalty term or loss
  - λ_{ik}: the loss incurred for classifying $\mathbf{x}$ into C_i when it belongs to C_k
    - Some wrong decisions have more serious implications than others
  - The loss matrix
    - $L = (\lambda_{ik})$
- Bayesian decision rule
  - To minimize the posterior expected risk
    - $\mathbf{x} \in R_i$ if $i = \underset{k}{\operatorname{argmin}}\, R(C_k|\mathbf{x})$
    - $R(C_i|\mathbf{x})$: the conditional risk when classifying $\mathbf{x}$ into C_i
      - $R(C_i|\mathbf{x}) = \sum_{k=1}^{K} \lambda_{ik} P(C_k|\mathbf{x}) = \lambda_{i1}P(C_1|\mathbf{x}) + \cdots + \lambda_{iK}P(C_K|\mathbf{x})$
- The overall risk
  - $R = \sum_{i=1}^{K} \int_{R_i} R(C_i|\mathbf{x})p(\mathbf{x})d\mathbf{x} = \sum_{i=1}^{K} \int_{R_i} \left(\sum_{k=1}^{K} \lambda_{ik}P(C_k|\mathbf{x})p(\mathbf{x})\right)d\mathbf{x}$

The overall risk is minimized if each of the integrals is minimized

Losses and Risks (2)

- Minimum-risk decision rule (for two-category case)
  - The conditional risk
    - $R(C_1|\mathbf{x}) = \lambda_{11}P(C_1|\mathbf{x}) + \lambda_{12}P(C_2|\mathbf{x})$
    - $R(C_2|\mathbf{x}) = \lambda_{21}P(C_1|\mathbf{x}) + \lambda_{22}P(C_2|\mathbf{x})$
  - The decision rule
    - $\mathbf{x} \to C_1$ if $R(C_1|\mathbf{x}) < R(C_2|\mathbf{x})$
    - $\mathbf{x} \to C_1$ if $\lambda_{11}P(C_1|\mathbf{x}) + \lambda_{12}P(C_2|\mathbf{x}) < \lambda_{21}P(C_1|\mathbf{x}) + \lambda_{22}P(C_2|\mathbf{x})$
    - $\mathbf{x} \to C_1$ if $(\lambda_{21} - \lambda_{11})P(C_1|\mathbf{x}) > (\lambda_{12} - \lambda_{22})P(C_2|\mathbf{x})$
    - $\mathbf{x} \to C_1$ if $(\lambda_{21} - \lambda_{11})p(\mathbf{x}|C_1)P(C_1) > (\lambda_{12} - \lambda_{22})p(\mathbf{x}|C_1)P(C_2)$
  - Assume that the loss incurred for making an error > the loss incurred for being correct
    - i.e., $(\lambda_{21} - \lambda_{11}) > 0, (\lambda_{12} - \lambda_{22}) > 0$
    - $\mathbf{x} \to C_1$ if $\boxed{\frac{p(\mathbf{x}|C_1)}{p(\mathbf{x}|C_2)}} > \boxed{\frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}}\frac{P(C_2)}{P(C_1)}}$

likelihood ratio

a threshold that is independent of the observation $\mathbf{x}$

if the threshold =1
=> Maximum likelihood decision rule

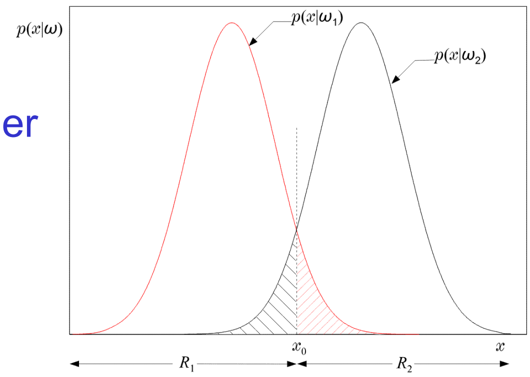
Losses and Risks (3)

- Example 2.1 [Theodoridis 09]
  - 2-class problem, with 1-D feature
    - $P(C_1) = P(C_2) = 0.5$, $\lambda_{12} > \lambda_{21}$
  - Assume
    - $L = \begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{bmatrix} = \begin{bmatrix} 0 & \lambda_{12} \\ \lambda_{21} & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0.5 & 0 \end{bmatrix}$
  - The minimum risk classifier is
    - $x \to C_1$ if $\frac{p(x|C_1)}{p(x|C_2)} > \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}}\frac{P(C_2)}{P(C_1)} = \frac{\lambda_{12}}{\lambda_{21}} = 2$
  - If the class-conditional probability density functions are
    - $p(x|C_1) = \frac{1}{\sqrt{\pi}}\exp(-x^2) \sim N(0, \frac{1}{2})$
    - $p(x|C_2) = \frac{1}{\sqrt{\pi}}\exp(-(x-1)^2) \sim N(1, \frac{1}{2})$

Losses and Risks (4)

- Example 2.1 (cont.)
  - The minimum probability error classifier
    - $x \to C_1$ if $P(C_1|x) > P(C_2|x)$
    - $x \to C_1$ if $p(x|C_1) > p(x|C_2)$
    - $x \to C_1$ if $\exp(-x^2) > \exp(-(x-1)^2)$
    - $x \to C_1$ if $x < \frac{1}{2}$
  - The minimum risk classifier
    - $x \to C_1$ if $\frac{p(x|C_1)}{p(x|C_2)} > \frac{\lambda_{12}}{\lambda_{21}} = 2$
    - $x \to C_1$ if $\exp(-x^2) > 2\exp(-(x-1)^2)$
    - $x \to C_1$ if $x < \frac{1-\ln 2}{2}$

Expanding the region R_2



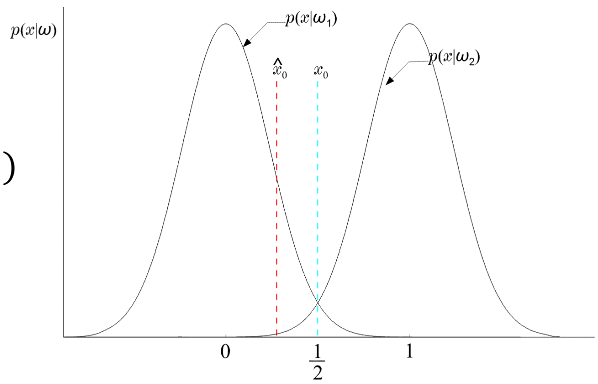


Fig. 2.1 [Theodoridis 09]

Losses and Risks (5)

- Special case
  - If all errors are equally costly
    - Zero-one loss function (0/1 loss)
    - $\lambda_{ik} = \begin{cases} 0, & i = k \\ 1, & i \neq k \end{cases}$
      - $\lambda_{ik} = \lambda_{ki}$
      - $\lambda_{ii} = 0$
  - Then
    - The conditional risk is simplified as
      - $R(C_i|\mathbf{x}) = \sum_{k=1}^{K} \lambda_{ik} P(C_k|\mathbf{x}) = \sum_{k \neq i} P(C_k|\mathbf{x}) = 1 - P(C_k|\mathbf{x})$
    - (minimizing the risk) = (minimizing the probability of error) = (maximizing the posterior probability)

Losses and Risks (6)

- Example [Duda, 01]
  - Case 1: $L = \begin{bmatrix} 0 & \lambda_{12} \\ \lambda_{21} & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \theta_a = \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}} \frac{P(C_2)}{P(C_1)} = \frac{P(C_2)}{P(C_1)}$
  - Case 2: $L = \begin{bmatrix} 0 & \lambda_{12} \\ \lambda_{21} & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1.2 \\ 1 & 0 \end{bmatrix}, \theta_b = \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}} \frac{P(C_2)}{P(C_1)} = 1.2 \frac{P(C_2)}{P(C_1)}$

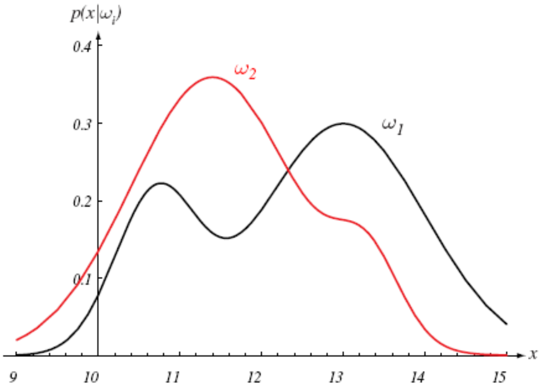


Fig. 2.1 [Duda 01]

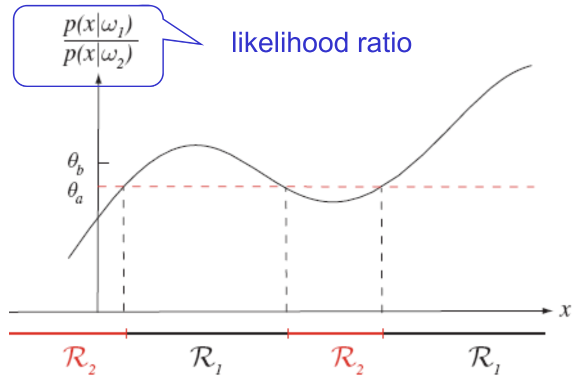


Fig. 2.3 [Duda 01]

Discriminant Functions (1)

- Minimizing either the overall risk or the error probability =>
  - Partitioning the feature space into K decision regions $R_1, \dots, R_K$
    - If the regions are contiguous, then they are separated by a decision surface $g_{ij}(\mathbf{x}) \equiv g_i(\mathbf{x}) - g_j(\mathbf{x}) = 0;\ \ i, j = 1, \dots, K;\ i \neq j$
  - Classifier
    - $\mathbf{x} \rightarrow C_i$ if $g_i(\mathbf{x}) > g_j(\mathbf{x})$, for all $j \neq i$
    - Discriminant functions
      - $g_i(\mathbf{x}), i = 1, \dots, K$

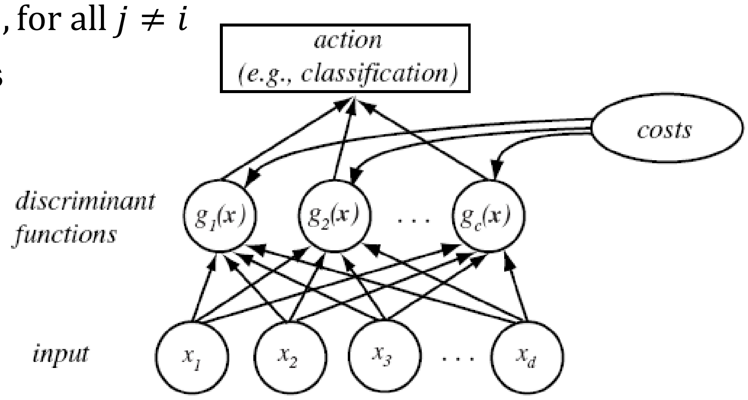


FIGURE 2.5. The functional structure of a general statistical pattern classifier which includes d inputs and c discriminant functions $g_i(\mathbf{x})$. A subsequent step determines which of the discriminant values is the maximum, and categorizes the input pattern

Fig. 2.5 [Duda 01]

Discriminant Functions (2)

- Discriminant function
  - In general, discriminant functions can be defined independent of the Bayes rule
    - Suboptimal solutions
  - Minimum-risk classifier
    - $g_i(\mathbf{x}) = -R(C_i|\mathbf{x})$
  - Minimum-error-rate classifier
    - $g_i(\mathbf{x}) = P(C_i|\mathbf{x})$
    - $g_i(\mathbf{x}) = p(\mathbf{x}|C_i)P(C_i)$
    - $g_i(\mathbf{x}) = \ln(p(\mathbf{x}|C_i)P(C_i)) = \ln p(\mathbf{x}|C_i) + \ln P(C_i)$
    - $g_i(\mathbf{x}) = f\big(P(C_i|\mathbf{x})\big)$
      - where $f(\cdot)$ is a monotonically increasing function

Discriminant Functions (3)

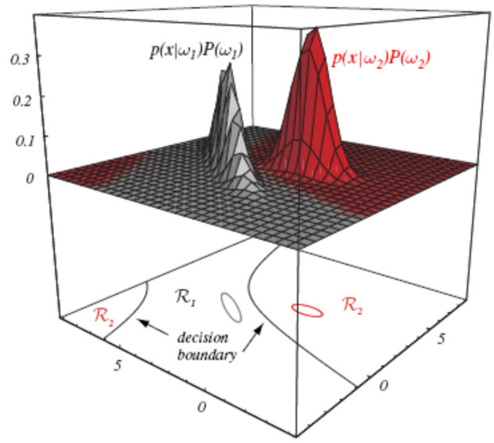


Fig. 2.6 [Duda 01]

- Two-category case
  - $\mathbf{x} \rightarrow C_1$ if $g_1(\mathbf{x}) > g_2(\mathbf{x})$
  - Or using a single discriminant function
    - $g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x})$
    - $\mathbf{x} \rightarrow C_1$ if $g(\mathbf{x}) > 0$
  - The minimum-error-rate classifier
    - $g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x}) = P(C_1|\mathbf{x}) - P(C_2|\mathbf{x})$
    - Or $g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x})$

$$= (\ln p(\mathbf{x}|C_1) + \ln P(C_1)) - (\ln p(\mathbf{x}|C_2) + \ln P(C_2))$$

$$= \ln\frac{p(\mathbf{x}|C_1)}{p(\mathbf{x}|C_2)} + \ln\frac{P(C_1)}{P(C_2)}$$

The Normal Density (1)

- Central limit theorem
  - The sum of a large number of independent, identically distributed random variables approximately follows a Gaussian distribution
- Univariate normal (Gaussian) density
  - The bell-shaped distribution
    - $p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
  - Is completely specified by its
    - Mean
      - $\mu = E\{x\} \equiv \int_{-\infty}^{\infty} x p(x) dx$
    - Variance
      - $\sigma^2 = E\{(x-\mu)^2\} \equiv \int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx$

$X \sim N(\mu, \sigma^2)$ denotes $p(X = x) = N(x|\mu, \sigma^2)$

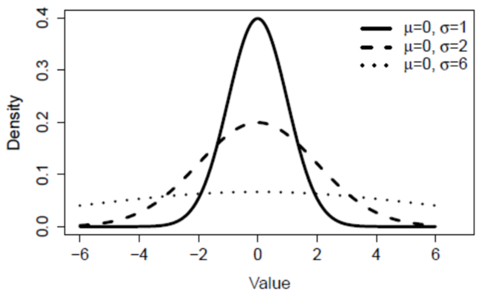


Figure: Three normal distributions with identical means but different standard deviations.

The Normal Density (2)

- Univariate normal density
  - The 68-95-99.7 rule
    - Approximately 68% of the values: within one σ of μ
    - Approximately 95% of the values: within two σ of μ
    - Approximately 99.7% of the values: within three σ of μ

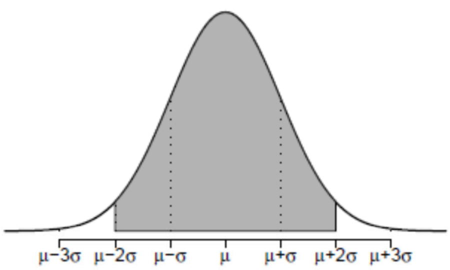


Figure: An illustration of the 68 – 95 – 99.7 percentage rule that a normal distribution defines as the expected distribution of observations. The grey region defines the area where 95% of observations are expected.

The Normal Density (3)

- Multivariate normal density $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathbf{x} = [x_1, \dots, x_l]^T$
  - $p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{l}{2}}\sqrt{|\boldsymbol{\Sigma}|}} \exp\left(-\frac{(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{2}\right)$

#parameters: $l + \frac{l(l+1)}{2}$

  - Mean vector
    - $\boldsymbol{\mu} = E\{\mathbf{x}\} = \int \mathbf{x}p(\mathbf{x})d\mathbf{x} = [\mu_1, \dots, \mu_l]^T$, $\mu_i = E\{x_i\}$
  - Covariance matrix
    - $\boldsymbol{\Sigma} = \mathrm{cov}[\mathbf{x}] = E\{(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\} = \int (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T p(\mathbf{x})d\mathbf{x}$

$$= \begin{bmatrix} \sigma_1^2 & \sigma_{12} \cdots & \sigma_{1l} \\ \sigma_{21} & \sigma_2^2 \cdots & \sigma_{2l} \\ \vdots & \ddots & \vdots \\ \sigma_{l1} & \sigma_{l2} \cdots & \sigma_l^2 \end{bmatrix}$$

- $\sigma_i^2 = E\{(x_i - \mu_i)^2\}$ variance of x_i
- $\sigma_{ij} = \sigma_{ji} = E\{(x_i - \mu_i)(x_j - \mu_j)\}$ covariance between x_i and x_j

The Normal Density (4)

- Multivariate normal density
  - Samples drawn from a normal density tend to fall in a single cloud
    - Cloud center: determined by the mean vector
    - Cloud shape: determined by the covariance matrix
      - The principal axes of hyperellipsoids are the eigenvectors of the covariance matrix

Eigen-decomposition of $\boldsymbol{\Sigma}$

$$\boldsymbol{\Sigma} = \boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}^T = \boldsymbol{\Phi}\begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}\boldsymbol{\Phi}^T$$

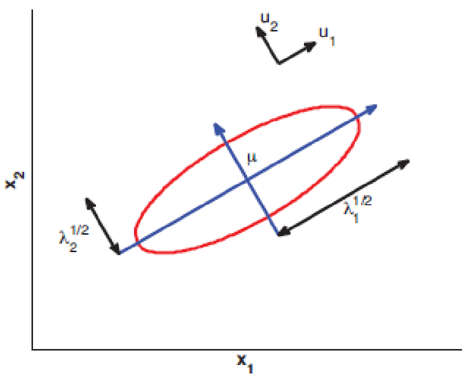


Fig. 4.1 [Murphy 2012]

Figure 4.1 Visualization of a 2 dimensional Gaussian density. The major and minor axes of the ellipse are defined by the first two eigenvectors of the covariance matrix, namely $\mathbf{u}_1$ and $\mathbf{u}_2$. Based on Figure 2.7 of (Bishop 2006a).

The Normal Density (5)

- Multivariate normal density
  - Any uncorrelated Gaussian random variables are also independent
    - This property is NOT shared by other distributions
  - Example ($l = 2$)
    - $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$
    - $p(\mathbf{x}) = p_{X_1,X_2}(x_1,x_2) = \frac{1}{(2\pi)\sigma_1\sigma_2}\exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}-\frac{(x_2-\mu_2)^2}{2\sigma_2^2}\right)$

      $= \frac{1}{\sigma_1\sqrt{2\pi}}\exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}\right)\frac{1}{\sigma_2\sqrt{2\pi}}\exp\left(-\frac{(x_2-\mu_2)^2}{2\sigma_2^2}\right) = p_{X_1}(x_1)p_{X_2}(x_2)$
    - $p(\mathbf{x})$ reduces to the product of the independent univariate normal densities $p(x_i)$

The Normal Density (6)

- Example (Bivariate Gaussian Density)
  - Correlated r.v.s - Isotropic uncorrelated r.v.s

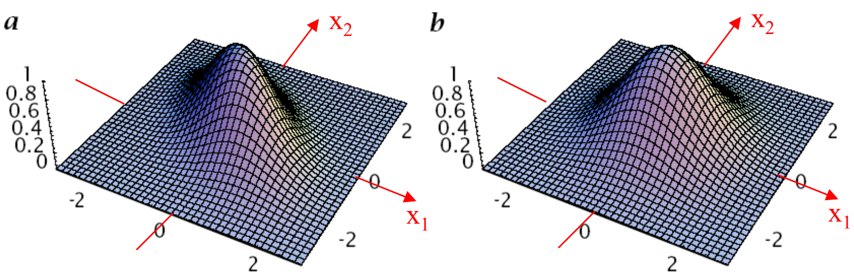


Fig. 3.4 [B. Jahne 02]

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix} = \boldsymbol{\Phi\Lambda\Phi}^T = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}\begin{bmatrix} 1.5 & 0 \\ 0 & 0.5 \end{bmatrix}\begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Eigen-decomposition of Σ

The Normal Density (7)

Figs. 2.3-2.6 [Theodoridis 09]

Spherical covariance matrix: circular shape

$$\boldsymbol{\Sigma} = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$$

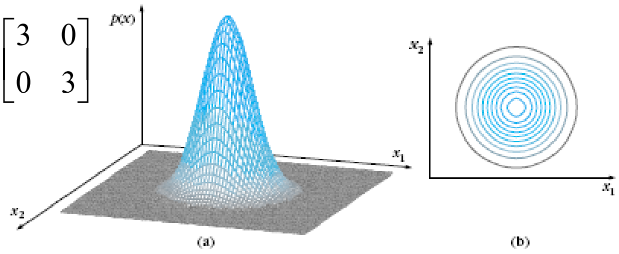


FIGURE 2.3

(a) The graph of a two-dimensional Gaussian pdf and (b) the corresponding isovalue curves for a diagonal Σ with $\sigma_1^2 = \sigma_2^2$. The graph has a spherical symmetry showing no preference in any direction.

Diagonal covariance matrix: axis-aligned ellipse

$$\boldsymbol{\Sigma} = \begin{bmatrix} 3 & 0 \\ 0 & 15 \end{bmatrix}$$

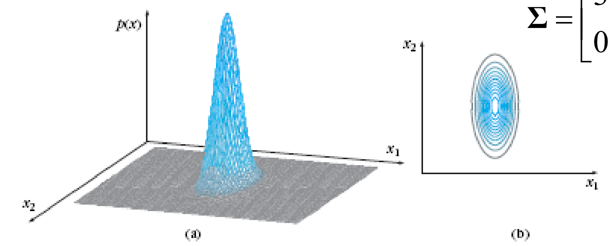


FIGURE 2.5

(a) The graph of a two-dimensional Gaussian pdf and (b) the corresponding isovalue curves for a diagonal Σ with $\sigma_1^2 << \sigma_2^2$. The graph is elongated along the x_2 direction.

Diagonal covariance matrix: axis-aligned ellipse

$$\boldsymbol{\Sigma} = \begin{bmatrix} 15 & 0 \\ 0 & 3 \end{bmatrix}$$

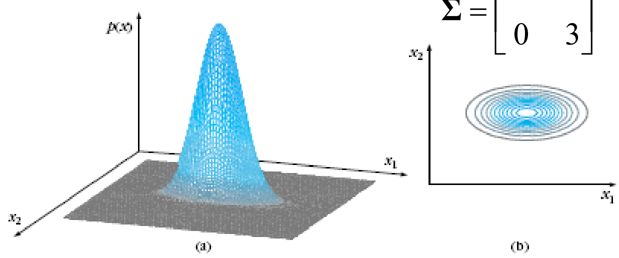


FIGURE 2.4

(a) The graph of a two-dimensional Gaussian pdf and (b) the corresponding isovalue curves for a diagonal Σ with $\sigma_1^2 >> \sigma_2^2$. The graph is elongated along the x_1 direction.

Full covariance matrix: elliptical contour

$$\boldsymbol{\Sigma} = \begin{bmatrix} 15 & 6 \\ 6 & 3 \end{bmatrix}$$

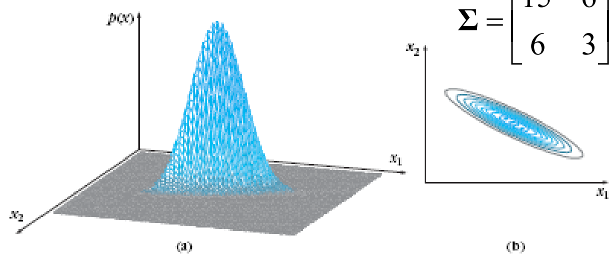


FIGURE 2.6

(a) The graph of a two-dimensional Gaussian pdf and (b) the corresponding isovalue curves for a case of a nondiagonal Σ. Playing with the values of the elements of Σ one can achieve different shapes and orientations.

The Normal Density (8)

- Linear transformation of random variables
  - If $\mathbf{x}$ is an l-dimensional random vector and $\mathbf{y} = \mathbf{A}^T\mathbf{x}$
    - $\mathbf{A}$ is a $l \times k$ matrix
    - $\mathbf{y}$ is a k-dimensional random vector
  - Then
    - We can easily derive the mean and covariance of $\mathbf{y}$
    - $\boldsymbol{\mu}_\mathbf{y} = E\{\mathbf{y}\} = E\{\mathbf{A}^T\mathbf{x}\} = \mathbf{A}^T\boldsymbol{\mu}_\mathbf{x}$
    - $\boldsymbol{\Sigma}_\mathbf{y} = E\left\{(\mathbf{y} - \boldsymbol{\mu}_\mathbf{y})(\mathbf{y} - \boldsymbol{\mu}_\mathbf{y})^T\right\} = E\left\{(\mathbf{A}^T\mathbf{x} - \mathbf{A}^T\boldsymbol{\mu}_\mathbf{y})(\mathbf{A}^T\mathbf{x} - \mathbf{A}^T\boldsymbol{\mu}_\mathbf{y})^T\right\}$
      $= E\{\mathbf{A}^T(\mathbf{x} - \boldsymbol{\mu}_\mathbf{x})(\mathbf{x} - \boldsymbol{\mu}_\mathbf{x})^T\} = \mathbf{A}^T E\{(\mathbf{x} - \boldsymbol{\mu}_\mathbf{x})(\mathbf{x} - \boldsymbol{\mu}_\mathbf{x})^T\}\mathbf{A}$
      $= \mathbf{A}^T\boldsymbol{\Sigma}_\mathbf{x}\mathbf{A}$
  - However, the mean and covariance only completely define the distribution of $\mathbf{y}$ if $\mathbf{x}$ is Gaussian
    - If $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \mathbf{y} = \mathbf{A}^T\mathbf{x}$ then $\mathbf{y} \sim N(\mathbf{A}^T\boldsymbol{\mu}, \mathbf{A}^T\boldsymbol{\Sigma}\mathbf{A})$

The Normal Density (9)

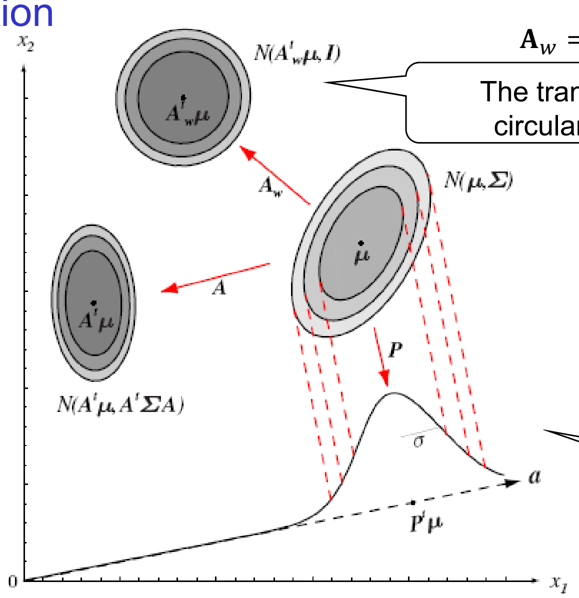
- Whitening transform
  - The transformed distribution has covariance matrix = identity matrix
    - The symmetric matrix $\mathbf{\Sigma}$ can be diagonalized by
      - $\mathbf{\Phi}^T\mathbf{\Sigma}\mathbf{\Phi} = \mathbf{\Lambda}$
      - $\mathbf{\Phi}$ is an orthogonal matrix having its columns the unit eigenvectors of $\mathbf{\Sigma}$
        - » $\mathbf{\Phi} = [\mathbf{v}_1 \quad \mathbf{v}_2 \quad \dots \quad \mathbf{v}_l]$
      - Λ is the diagonal matrix containing the corresponding eigenvalues of $\mathbf{\Sigma}$
        - » $\mathbf{\Lambda} = diag(\lambda_1, \lambda_2, \dots, \lambda_l)$
    - Then with the transform
      - $\mathbf{A}_w = \mathbf{\Phi}\mathbf{\Lambda}^{-\frac{1}{2}} \Rightarrow \mathbf{A}_w^T\mathbf{\Sigma}\mathbf{A}_w = \mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{\Phi}^T\mathbf{\Sigma}\mathbf{\Phi}\mathbf{\Lambda}^{-\frac{1}{2}} = \mathbf{I}$
      - If $\mathbf{x} \sim N(\boldsymbol{\mu}, \mathbf{\Sigma}), \mathbf{y} = \mathbf{A}_w^T\mathbf{x}$
        - » $\mathbf{y} \sim N\left(\mathbf{A}_w^T\boldsymbol{\mu}, \mathbf{A}_w^T\mathbf{\Sigma}\mathbf{A}_w\right) = N(\mathbf{A}_w^T\boldsymbol{\mu}, \mathbf{I}\ )$
      - The product of l independent univariate Gaussian distributions

The Normal Density (10)

- Example
  - The action of a linear transformation on the feature space will convert an arbitrary normal distribution into another normal distribution

$\mathbf{A}_w = \mathbf{\Phi}\mathbf{\Lambda}^{-\frac{1}{2}}$

The transformed distribution is a circularly symmetric Gaussian



$\mathbf{P} = [\mathbf{a}]_{l\times 1}$

The transform projects $\mathbf{x}$ onto a line in the direction of $\mathbf{a}$

Fig. 2.8 [Duda 01]

Bayesian Classification For Normal Distribution (1)

- Goal
  - To study the optimal Bayesian classifier when the involved pdfs $p(\mathbf{x}|C_i), i = 1, \ldots, K$ are multivariate normal distributions
- Discriminant function of the minimum-error-rate classifier
  - $g_i(\mathbf{x}) = \ln p(\mathbf{x}|C_i) + \ln P(C_i)$
  - Assume the likelihood functions of C_i w.r.t. $\mathbf{x}$ in the l-dimensional feature space follow the multivariate normal density
    - $p(\mathbf{x}|C_i) = N(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{\frac{l}{2}}\sqrt{|\boldsymbol{\Sigma}_i|}} \exp(-\frac{(\mathbf{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)}{2})$
  - Then
    - $g_i(\mathbf{x}) = -\frac{(\mathbf{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)}{2} - \frac{l}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(C_i)$

Bayesian Classification For Normal Distribution (2)

- Case 1: $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma} = \sigma^2\mathbf{I}$ (isotropic covariance)
  - Assume the covariance matrix is the same in all classes
  - Assume the features x_k are statistically independent and each has the same variance σ^2
    - $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma} = \sigma^2\mathbf{I}$, $|\boldsymbol{\Sigma}_i| = \sigma^{2l}, \boldsymbol{\Sigma}_i^{-1} = (1/\sigma^2)\mathbf{I}$
  - Ignoring the terms independent of i
    - $g_i(\mathbf{x}) = -\frac{(\mathbf{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)}{2} + \ln P(C_i) = -\frac{(\mathbf{x}-\boldsymbol{\mu}_i)^T(\mathbf{x}-\boldsymbol{\mu}_i)}{2\sigma^2} + \ln P(C_i)$
      $= -\frac{\|\mathbf{x}-\boldsymbol{\mu}_i\|^2}{2\sigma^2} + \ln P(C_i) = -\frac{1}{2\sigma^2}\left(\mathbf{x}^\mathrm{T}\mathbf{x} - 2\boldsymbol{\mu}_i^T\mathbf{x} + \boldsymbol{\mu}_i^T\boldsymbol{\mu}_i\right) + \ln P(C_i)$
  - Ignoring the term $\mathbf{x}^\mathrm{T}\mathbf{x}$ which is the same for all i
    - $g_i(\mathbf{x}) = \left(\frac{1}{\sigma^2}\boldsymbol{\mu}_i^T\right)\mathbf{x} + \left(-\frac{1}{2\sigma^2}\boldsymbol{\mu}_i^T\boldsymbol{\mu}_i + \ln P(C_i)\right) = \mathbf{w}_i^T\mathbf{x} + w_{i0}$

Linear discriminant function

Bayesian Classification For Normal Distribution (3)

- Case 1 (cont.): $\mathbf{\Sigma}_i = \mathbf{\Sigma} = \sigma^2\mathbf{I}$
  - $g_i(\mathbf{x}) = \mathbf{w}_i^T\mathbf{x} + w_{i0}$ is a linear function of $\mathbf{x}$
    - The decision surfaces are hyperplanes defined by
      - $g_{ij}(\mathbf{x}) \equiv g_i(\mathbf{x}) - g_j(\mathbf{x}) = 0$
      - $\mathbf{w}^T(\mathbf{x} - \mathbf{x_0}) = 0$ — The hyperplane passes through $\mathbf{x_0}$
      - Where
        - » $\mathbf{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$
        - » $\mathbf{x_0} = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\sigma^2}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2} \ln\frac{P(C_i)}{P(C_j)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$
    - If $P(C_i) \neq P(C_j)$
      - The point $\mathbf{x_0}$ shifts away from the more likely mean
    - If $P(C_i) = P(C_j)$
      - $\mathbf{x_0} = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j)$
      - The point $\mathbf{x_0}$ is halfway between the means and the hyperplane is the perpendicular bisector of the line between the means

Bayesian Classification For Normal Distribution (4)

- Case 1 (cont.): $\mathbf{\Sigma}_i = \mathbf{\Sigma} = \sigma^2\mathbf{I}$
  - Decision boundary
    - $\mathbf{w}^T(\mathbf{x} - \mathbf{x_0}) = 0$
    - The hyperplane is orthogonal to the line linking the means

$\because \mathbf{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$

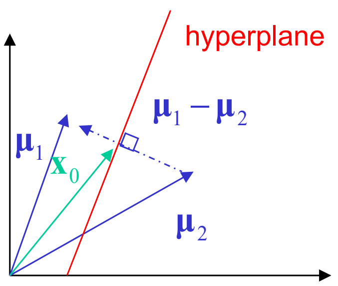
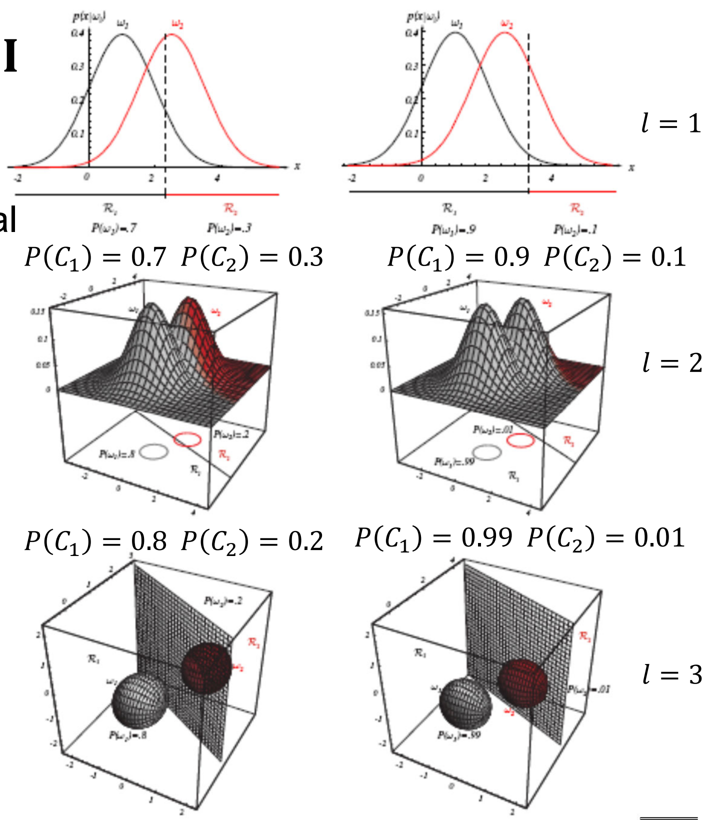


Fig. 2.11 [Duda 01]



Bayesian Classification For Normal Distribution (5)

- Case 1 (cont.): $\mathbf{\Sigma}_i = \mathbf{\Sigma} = \sigma^2\mathbf{I}$
  - If the prior probabilities are the same for all K classes
    - (i.e. equiprobable classes with the same covariance matrix)
    - $g_i(\mathbf{x}) = -\|\mathbf{x} - \boldsymbol{\mu}_i\|^2$ — Minimum-distance classifier
    - Maximum $g_i(\mathbf{x})$ = Minimum the Euclidean distance $\|\mathbf{x} - \boldsymbol{\mu}_i\|^2$
    - Feature vectors are assigned to classes of the nearest mean

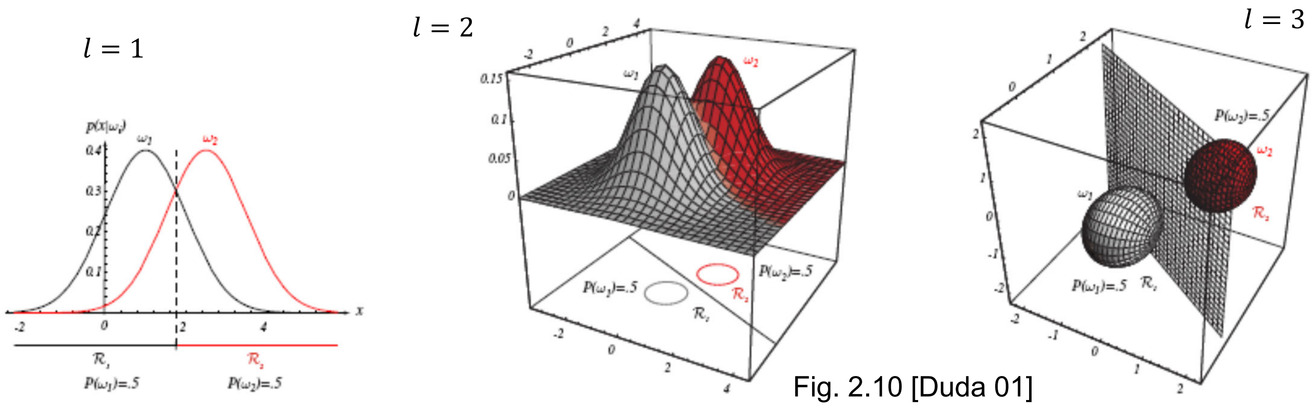


Fig. 2.10 [Duda 01]

Bayesian Classification For Normal Distribution (6)

- Case 1 (cont.): $\mathbf{\Sigma}_i = \mathbf{\Sigma} = \sigma^2\mathbf{I}$

$P(C_i) < P(C_j)$　　$P(C_i) = P(C_j)$

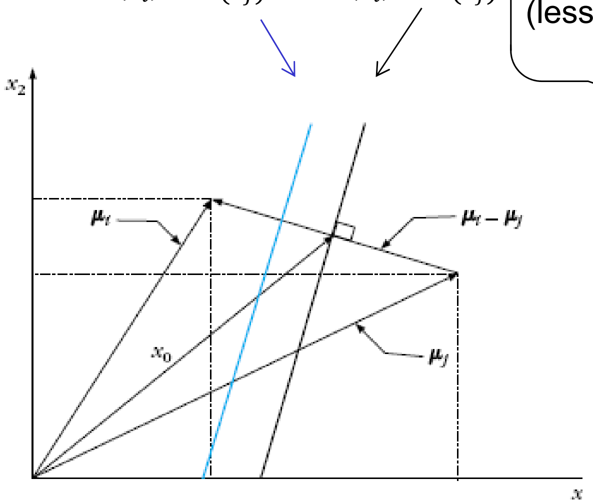


Fig. 2.10 [Theodoridis 09]

Decision line for compact classes (less sensitive to the values of $P(C_i), P(C_j)$)

Decision line for noncompact classes (a small movement of hyperplane may be critical)

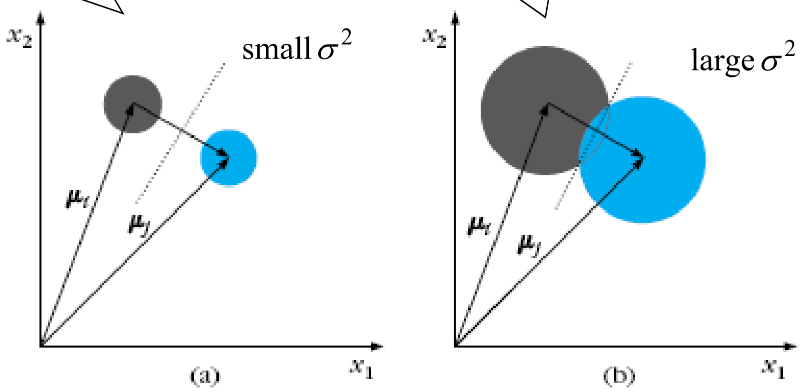


Fig. 2.11 [Theodoridis 09]

Bayesian Classification For Normal Distribution (7)

- Case 2: $\mathbf{\Sigma}_i = \mathbf{\Sigma} = diag(\sigma_1^2, \sigma_2^2, \dots, \sigma_l^2)$
  - Assume the features x_j are statistically independent but may have different variance
  - Classes are hyperellipsodial and axis-aligned
    - $g_i(\mathbf{x}) = -\frac{(\mathbf{x}-\boldsymbol{\mu}_i)^T \mathbf{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)}{2} + \ln P(C_i)$

      $= -\frac{1}{2}\sum_{j=1}^{l}\left(\frac{x_j-\mu_{ij}}{\sigma_j}\right)^2 + \ln P(C_i)$
    - $g'_i(\mathbf{x}) = \sum_{j=1}^{l}\left(\frac{\mu_{ij}}{\sigma_j^2}\right)x_j + \left(-\frac{1}{2}\sum_{j=1}^{l}\left(\frac{\mu_{ij}^2}{\sigma_j^2}\right) + \ln P(C_i)\right)$

      $= \mathbf{w}_i^T \mathbf{x} + w_{i0}$

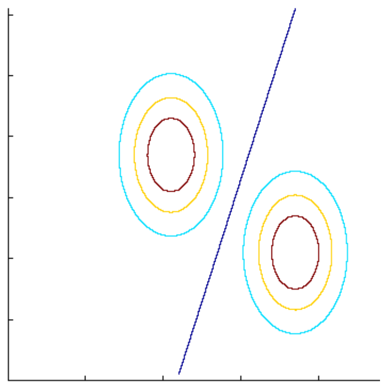


Fig. 5.5 [Alpaydin, 2014]

Bayesian Classification For Normal Distribution (8)

- Case 3: $\mathbf{\Sigma}_i = \mathbf{\Sigma}$
  - The covariance matrices for all classes are the same but otherwise arbitrary
    - $g_i(\mathbf{x}) = -\frac{(\mathbf{x}-\boldsymbol{\mu}_i)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)}{2} + \ln P(C_i)$

      $= -\frac{1}{2}\left[\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x} - 2\boldsymbol{\mu}_i^T\mathbf{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}_i^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_i\right] + \ln P(C_i)$
    - Ignoring the terms independent of i
    - Linear discriminant function => the decision surfaces are hyperplanes
    - $g_i(\mathbf{x}) = \mathbf{w}_i^T \mathbf{x} + w_{i0}$
      - $\mathbf{w}_i = \mathbf{\Sigma}^{-1}\boldsymbol{\mu}_i \quad w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln P(C_i)$
  - If the prior probabilities are the same for all K classes
    - $g_i(\mathbf{x}) = -\frac{(\mathbf{x}-\boldsymbol{\mu}_i)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)}{2}$
    - Maximum $g_i(\mathbf{x})$ = Minimum the Mahalanobis distance from $\mathbf{x}$ to $\boldsymbol{\mu}_i$

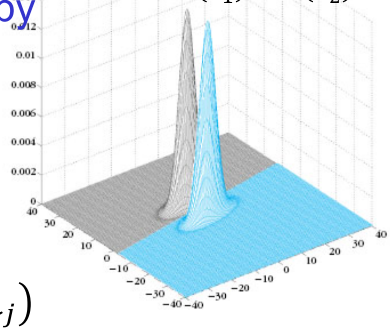
      $\left((\mathbf{x}-\boldsymbol{\mu}_i)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right)^{\frac{1}{2}}$

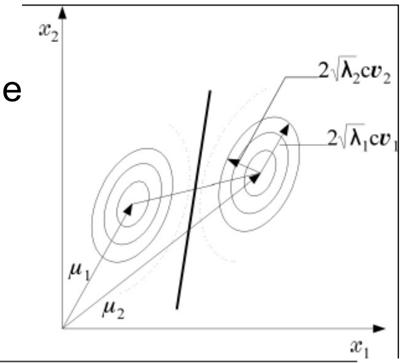
Bayesian Classification For Normal Distribution (9)

Fig. 2.12 & Fig. 2.13(b) [Theodoridis 09]

- Case 3 (cont.): $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$
  - The decision surfaces are hyperplanes defined by
    - $g_{ij}(\mathbf{x}) \equiv g_i(\mathbf{x}) - g_j(\mathbf{x}) = 0$
    - $\Rightarrow \mathbf{w}^T(\mathbf{x} - \mathbf{x_0}) = 0$
    - $\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$
    - $\mathbf{x_0} = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{1}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)} \ln \frac{P(C_i)}{P(C_j)} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$
    - The hyperplane is generally NOT orthogonal to the line between the means but to its linear transform
      - $\because \mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$

$P(C_1) = P(C_2)$

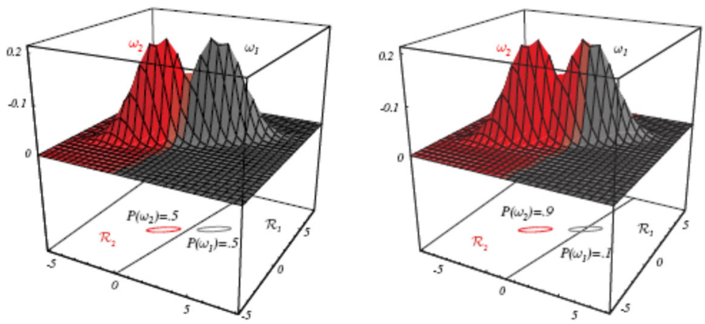




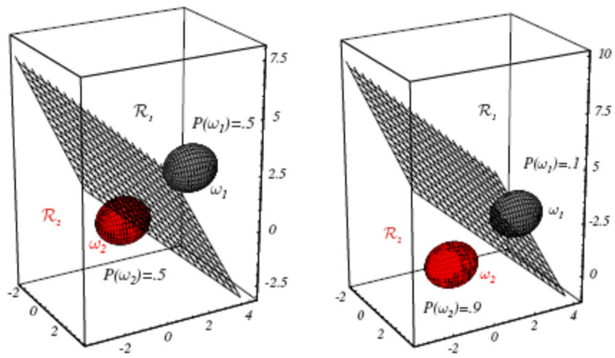
Bayesian Classification For Normal Distribution (10)

- Case 3 (cont.): $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$
  - Decision boundary

Fig. 2.12 [Duda 01]

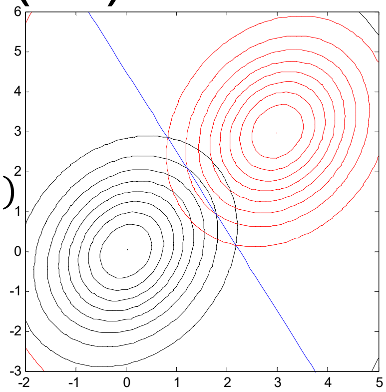


$P(C_1) = 0.5 \;\; P(C_2) = 0.5$ $P(C_1) = 0.1 \;\; P(C_2) = 0.9$ $l = 2$



$l = 3$

Bayesian Classification For Normal Distribution (11)



- Case 3 (cont.): $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$
  - Example 2.2 ($l = 2$) [Theodoridis 09]
    - Assume equal prior probabilities $P(C_1) = P(C_2)$
    - $K = 2$
      - $p(\mathbf{x}|C_1) \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}), \boldsymbol{\mu}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} 1.1 & 0.3 \\ 0.3 & 1.9 \end{bmatrix}$
      - $p(\mathbf{x}|C_2) \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}), \boldsymbol{\mu}_2 = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$
    - The Bayesian classifier = maximizing $g_i(\mathbf{x})$ = minimizing the Mahalanobis distance from $\mathbf{x}$ to $\boldsymbol{\mu}_i$
    - To classify a pattern $\mathbf{x} = [1, 2.2]^{\boldsymbol{T}}$
      - $d^2(\mathbf{x}, \boldsymbol{\mu}_1) = (\mathbf{x} - \boldsymbol{\mu}_1)^{\boldsymbol{T}} \boldsymbol{\Sigma}^{-\mathbf{1}} (\mathbf{x} - \boldsymbol{\mu}_1) = [1, 2.2] \begin{bmatrix} 0.95 & -0.15 \\ -0.15 & 0.55 \end{bmatrix} \begin{bmatrix} 1 \\ 2.2 \end{bmatrix} = 2.952$
      - $d^2(\mathbf{x}, \boldsymbol{\mu}_2) = [-2, -0.8] \begin{bmatrix} 0.95 & -0.15 \\ -0.15 & 0.55 \end{bmatrix} \begin{bmatrix} -2 \\ -0.8 \end{bmatrix} = 3.672$
      - $\therefore \mathbf{x} \to C_1$

Bayesian Classification For Normal Distribution (12)

- Case 4: $\boldsymbol{\Sigma}_i = \text{arbitrary}$
  - The covariance matrices are different for each category
    - The discriminant functions are nonlinear quadratic
    - $g_i(\mathbf{x}) = -\frac{(\mathbf{x}-\boldsymbol{\mu}_i)^{\boldsymbol{T}} \boldsymbol{\Sigma}_i^{-\mathbf{1}} (\mathbf{x}-\boldsymbol{\mu}_i)}{2} - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(C_i)$

$$= -\frac{1}{2}\left[\mathbf{x}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{x} - 2\boldsymbol{\mu}_i{}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{x} + \boldsymbol{\mu}_i{}^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\mu}_i\right] - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(C_i)$$

$$= \mathbf{x}^T (-\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}) \mathbf{x} + \left(\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i\right)^T \mathbf{x} + \left(-\frac{1}{2}\boldsymbol{\mu}_i{}^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\mu}_i - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(C_i)\right)$$

$$= \mathbf{x}^T \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^T \mathbf{x} + w_{i0}$$

  - Example ($l = 1$)
    - $P(C_1) = P(C_2)$
    - $\sigma_1^2 \neq \sigma_2^2$
    - $g(x) = g_1(x) - g_2(x) = ax^2 + bx + c$

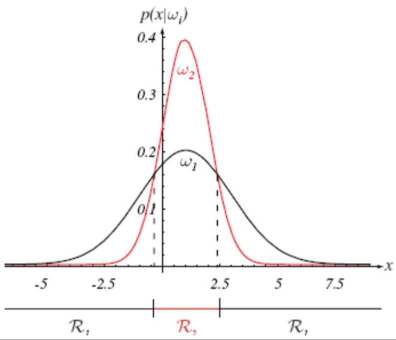


Fig. 2.13 [Duda 01]

Bayesian Classification For Normal Distribution (13)

- Case 4 (cont.)
  - Example ($l = 2$)

Decision boundaries are quadrics

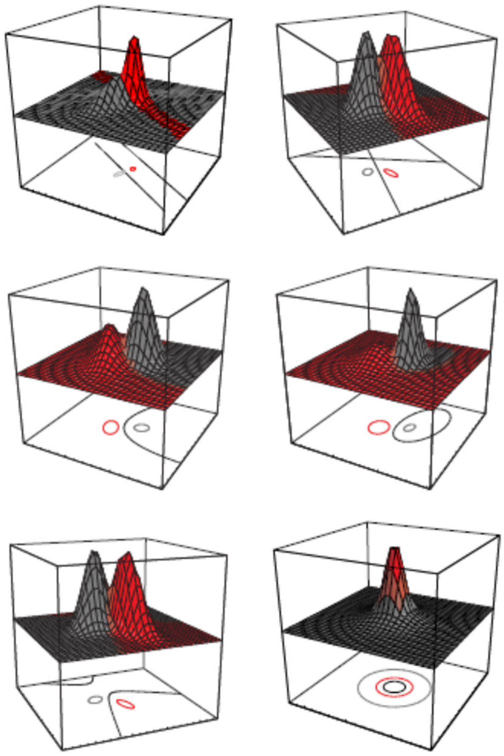
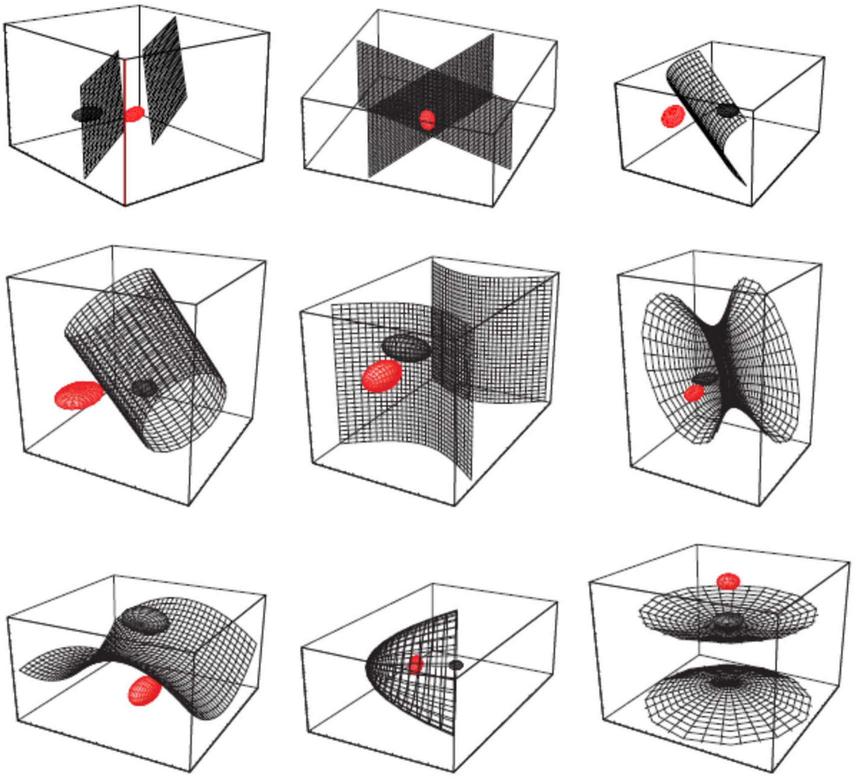


Fig. 2.14 [Duda 01]

Bayesian Classification For Normal Distribution (14)

- Case 4 (cont.)
  - Example ($l = 3$)

Decision boundaries are hyperquadrics



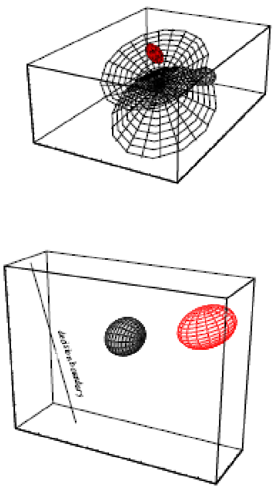
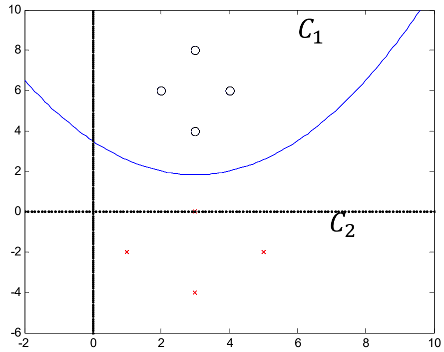


Fig. 2.15 [Duda 01]

Bayesian Classification For Normal Distribution (15)

- Case 4 (cont.)
  - Example (l = 2) [p.44, Duda 01]
    - Assume equal prior probabilities $P(C_1) = P(C_2)$
    - Let $K = 2$
      - $p(\mathbf{x}|C_1) \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \boldsymbol{\mu}_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}, \boldsymbol{\Sigma}_1 = \begin{bmatrix} 0.5 & 0 \\ 0 & 2 \end{bmatrix}, \boldsymbol{\Sigma}_1^{-1} = \begin{bmatrix} 2 & 0 \\ 0 & 0.5 \end{bmatrix}$
      - $p(\mathbf{x}|C_2) \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2), \boldsymbol{\mu}_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}, \boldsymbol{\Sigma}_2 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, \boldsymbol{\Sigma}_2^{-1} = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}$
      - $g_1(\mathbf{x}) = -\frac{1}{4}(4x_1^2 - 24x_1 + x_2^2 - 12x_2 + 72)$
      - $g_2(\mathbf{x}) = -\frac{1}{4}(x_1^2 - 6x_1 + x_2^2 + 4x_2 + 13) - \ln 2$
    - The decision boundary
      - $g_1(\mathbf{x}) - g_2(\mathbf{x}) = 0$
      - $x_2 = 0.1875(x_1 - 3)^2 + 1.83$

A parabola with vertex at $(3, 1.83)$



Bayesian Classification For Normal Distribution (16)

- Case 4 (cont.)
  - Example (l = 2) (p. 25, [Theodoridis 09])
    - Assume equal prior probabilities $P(C_1) = P(C_2)$
      - $p(\mathbf{x}|C_1) \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), p(\mathbf{x}|C_2) \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$

$\boldsymbol{\mu}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 0.3 & 0 \\ 0 & 0.35 \end{bmatrix}$

$\boldsymbol{\mu}_2 = \begin{bmatrix} 4 \\ 0 \end{bmatrix}$

$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 1.2 & 0 \\ 0 & 1.85 \end{bmatrix}$

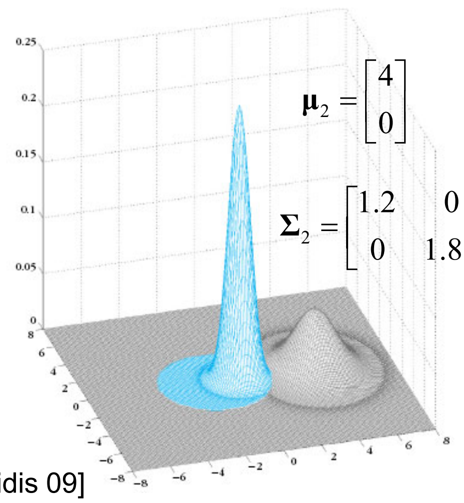


Fig. 2.8 [Theodoridis 09]

Elliptic decision curve

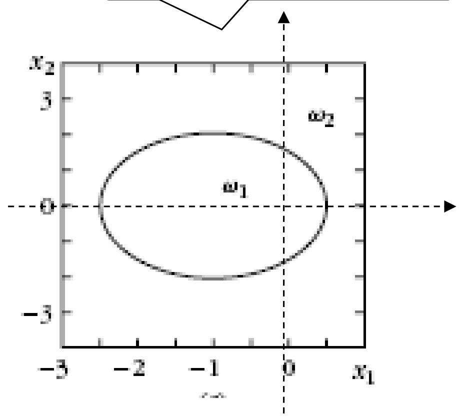


Fig. 2.7a [Theodoridis 09]

Bayesian Classification For Normal Distribution (17)

- Case 4 (cont.)
  - Example ($l = 2$) (p. 25, [Theodoridis 09])

Hyperbolic decision curve

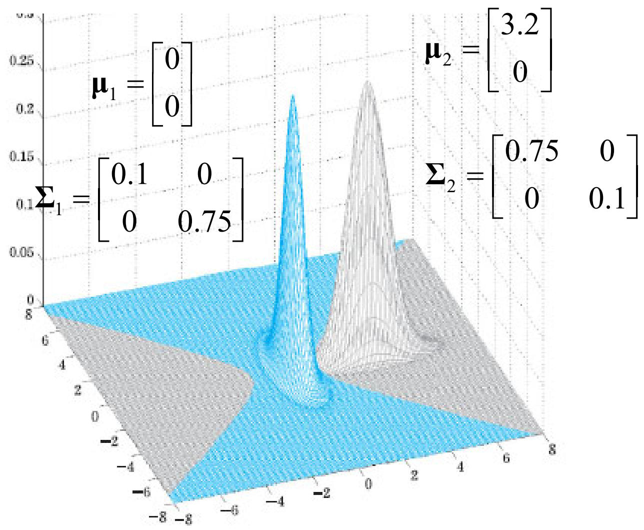


Fig. 2.9 [Theodoridis 09]

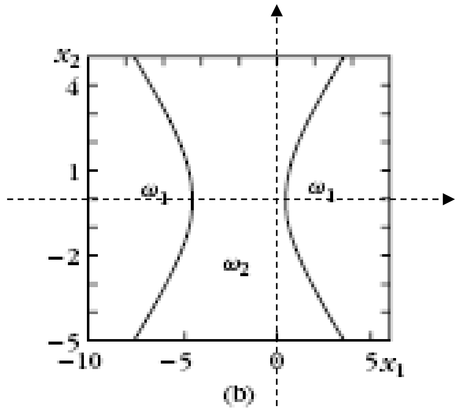


Fig. 2.7b [Theodoridis 09]

Naïve-Bayes Classifier (1)

- Naïve Bayes assumption
  - The distributions of the individual features are assumed to be conditional independent given the class label
    - $p(\mathbf{x}|C_i) = \prod_{j=1}^{l} p(x_j|C_i)$
    - To simplify the calculation of the full joint pdf $p(\boldsymbol{x})$
      - May suffers from the curse of dimensionality
  - The naïve Bayes classifier
    - $C_m = \underset{C_i}{\operatorname{argmax}} P(C_i) \prod_{j=1}^{l} p(x_j|C_i)$, $i = 1, 2, \ldots, K$

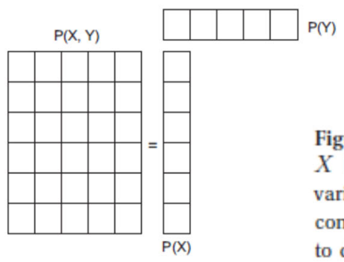


Figure 2.2 Computing $p(x, y) = p(x)p(y)$, where $X \perp Y$. Here X and Y are discrete random variables; X has 6 possible states (values) and Y has 5 possible states. A general joint distribution on two such variables would require $(6 \times 5) - 1 = 29$ parameters to define it (we subtract 1 because of the sum-to-one constraint). By assuming (unconditional) independence, we only need $(6 - 1) + (5 - 1) = 9$ parameters to define $p(x, y)$.

Fig. 2.2 [Murphy 2012]

Naïve-Bayes Classifier (2)

- Note, in the normal distribution cases
  - Uncorrelated Gaussian random variables are also independent
    - If $\Sigma = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_l^2)$
      - $p(\boldsymbol{x}|C_i)$ is reduced to the product of the independent univariate normal densities $p(x_j|C_i)$
  - Assume the features x_k are statistically independent but may have different variance
    - $\Sigma_i = diag(\sigma_{i1}^2, \sigma_{i2}^2, \ldots, \sigma_{il}^2)$
    - Equal covariance matrices for all the classes
      - Case 2: $\Sigma_i = \Sigma = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_l^2)$
    - Different covariance matrices
      - Special case of Case 4
        - » e.g., $\boldsymbol{\Sigma_1} = \begin{bmatrix} 0.5 & 0 \\ 0 & 2 \end{bmatrix}, \boldsymbol{\Sigma_2} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$

Naïve-Bayes Classifier (3)

- Example – binary discrete features
  - The feature vector $\boldsymbol{x} = [x_1, \ldots, x_l]^T$ with binary attributes $x_j \in \{0,1\}$
    - Let $p_{ij} \equiv P(x_j = 1|C_i)$
    - Adopting statistical independent assumption
      - $P(\boldsymbol{x}|C_i) = \prod_{j=1}^{l} P(x_j|C_i) = \prod_{j=1}^{l} p_{ij}^{x_j}(1-p_{ij})^{(1-x_j)}$
    - Then the discriminant function is a linear discriminant function
      - $g_i(\boldsymbol{x}) = \ln P(\boldsymbol{x}|C_i) + \ln P(C_i)$

$$= \sum_{j=1}^{l}[x_j \ln p_{ij} + (1-x_j)\ln(1-p_{ij})] + \ln P(C_i)$$

$$= \sum_{j=1}^{l}(x_j \ln \frac{p_{ij}}{1-p_{ij}}) + \sum_{j=1}^{l} \ln(1-p_{ij}) + \ln P(C_i)$$

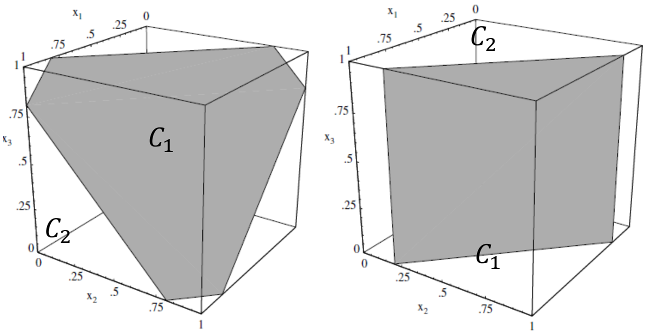
$$= \boldsymbol{w}_i^T \boldsymbol{x} + w_{i0}$$

- » $\boldsymbol{w}_i = [\ln \frac{p_{i1}}{1-p_{i1}}, \ldots, \ln \frac{p_{il}}{1-p_{il}}]^T$
- » $w_{i0} = \sum_{j=1}^{l} \ln(1-p_{ij}) + \ln P(C_i)$

Naïve-Bayes Classifier (4)

- Example [p.53, Duda, 01]
  - Consider a 2-class problem having 3 independent binary features with known feature probabilities $p_{ij},\ i = 1,2; j = 1,2,3$
    - If $P(C_1) = P(C_2)$
    - Case 1
      - $p_{11} = p_{12} = p_{13} = 0.8$
      - $p_{21} = p_{22} = p_{23} = 0.5$
      - $g(\boldsymbol{x}) = 1.3863(x_1 + x_2 + x_3) - 2.7489$
    - Case 2
      - $p_{11} = p_{12} = 0.8, p_{13} = 0.5$
      - $p_{21} = p_{22} = p_{23} = 0.5$
      - $g(\boldsymbol{x}) = 1.3863(x_1 + x_2) - 1.8326$

feature x_3 gives no predictive information about the categories



Naïve-Bayes Classifier (5)

- The disease classifier [Kelleher et al., 2015]
  - Assuming conditional independence between the 3 features
  - Given the patient with the measured features
    - $\boldsymbol{x}' = [headache = T, fever = T, vomiting = F]$
    - Priors
      - $P(C_M) = 0.3, P(\neg C_M) = 0.7$
    - Likelihoods
      - $P(\boldsymbol{x} = [T,T,F]|C_M) = \frac{2}{3} \times \frac{1}{3} \times \frac{1}{3}$
      - $P(\boldsymbol{x} = [T,T,F]|\neg C_M) = \frac{5}{7} \times \frac{3}{7} \times \frac{3}{7}$
    - The posterior probabilities
      - $P(C_M|\boldsymbol{x} = [T,T,F]\,) = 0.1948$
      - $P(\neg C_M|\boldsymbol{x} = [T,T,F]\,) = 0.8052$
    - The model is relatively robust to the curse of dimensionality
      - Especially important in scenarios with small datasets

ID	HEADACHE	FEVER	VOMITING	MENINGITIS
1	true	true	false	false
2	false	true	false	false
3	true	false	true	false
4	true	false	true	false
5	false	true	false	true
6	true	false	true	false
7	true	false	true	false
8	true	false	true	true
9	false	true	false	false
10	true	false	true	true